AN INTERPRETABLE MACHINE LEARNING FRAMEWORK FOR ACCURATE STROKE PREDICTION WITH A USER-FRIENDLY FRONTEND INTERFACE

Abstract

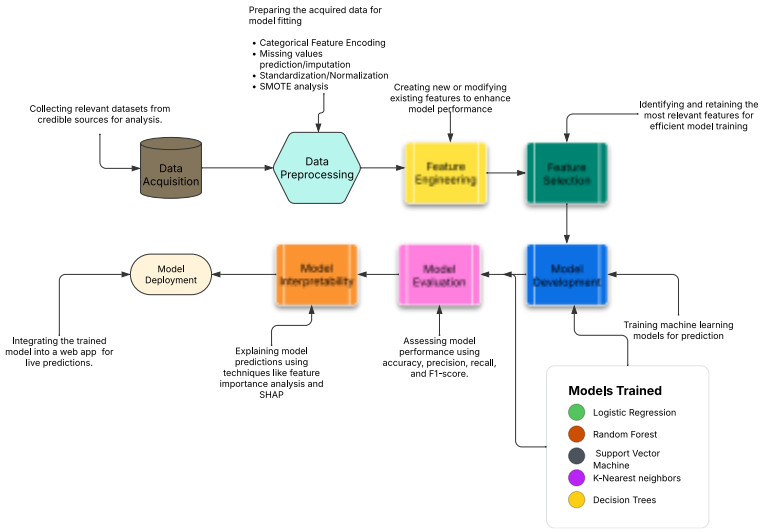
Stroke remains a leading cause of death and long-term disability worldwide. Late diagnosis typically results in permanent paralysis and even death. Early and accurate risk prediction has hence proven to be essential. With the advancement of artificial intelligence, machine learning (ML) provides a powerful means of stroke prediction based on historical patient data. However, model interpretability in healthcare is crucial for clinical acceptance and reliability. This research proposes an interpretable machine learning system for the prediction of stroke occurrence. A total of five models -Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbors, and Decision Tree - are trained and compared on critical evaluation metrics such as accuracy, precision, recall, and AUC-ROC scores. The Random Forest model results in the highest prediction accuracy and, hence, is identified as the most appropriate model. To complement transparency, SHapley Additive exPlanations (SHAP) are included to clarify feature importance and decision-making mechanisms. The final model is deployed as a web app via Streamlit, thus ensuring ease of access for researchers and healthcare professionals. Results illustrate that the proposed approach improves the accuracy of stroke prediction while maintaining interpretability, ultimately enhancing preventive healthcare measures.

Introduction

Stroke is a cerebrovascular disease marked by a sudden onset and its profound impact on affected individuals[1]. Each year, approximately fifteen million people worldwide experience a stroke, with five million suffering from significant disabilities as a result. Since 1990, data indicate a notable decline in stroke mortality rates in high-income countries; however, low- and middle-income countries (LMICs), including Ghana, have observed little to no substantial improvements[2]. Consequently, LMICs bear a greater burden from strokes due to factors such as population growth and insufficient stroke-related resources. Moreover, many people in Africa tend to overlook the early warning signs of a stroke, often believing that the symptoms will resolve on their own. Despite these challenges, Artificial Intelligence (AI) demonstrates significant potential to enhance healthcare outcomes, from mobile-based diagnostics to precision medicine. A delayed stroke diagnosis can lead to irreversible brain damage, highlighting the urgency of early risk prediction. A 2024 study indicates that the full integration of machine learning technologies into Ghana's healthcare system remains unrealized[3]. This gap offers a considerable opportunity for growth and development, as the advantages of machine learning could greatly improve health outcomes in the country. Therefore, this paper proposes the utilization of machine learning algorithms to predict stroke risk based on relevant risk factors.

Methodology

The dataset used in this study was sourced from Kaggle, comprising 5,110 patient observations across 11 distinct features, like patients' age, heart disease status, and BMI. The dataset underwent thorough examination and processing. An Exploratory Data Analysis and Preprocessing were conducted, during which missing values were predicted using a decision tree regressor pipeline. Correlation analysis was performed to visualize how each feature contributed to the risk of stroke. Feature engineering and selection were employed to optimize model performance, while class imbalance was addressed using SMOTE analysis. The data was divided into 80% for training and 20% for testing. Five models—Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbors, and Decision Tree—were trained and evaluated. The models were assessed based on accuracy, precision, recall, F1-score, and AUC-ROC. The figure below illustrates the methodology used in this project.



Stroke Prediction Methodology Figure

The obtained results are as follows:

Model	Accuracy	AUC-ROC score
Logistic Regression	78.5%,	87.0%
Random Forest: Accuracy	91.9%,	97.4%
Support Vector Machine	73.3%	
K-Nearest Neighbors	88.5%	
Decision Tree	85.9%	

Conclusion

This study highlights the effectiveness of machine learning in stroke prediction, with the Random Forest model achieving the highest accuracy. A Streamlit-based interface ensures accessibility for real-time risk assessments. Future efforts will focus on improving interpretability and incorporating additional clinical data for broader healthcare applications.

References

[1] World Health Organization, “Information resources Stroke, Cerebrovascular accident,” Oct. 2024. [Online]. Available: https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html

[2] Banerjee Tapas Kumar, “Global, Regional, and Country-Specific Lifetime Risks of Stroke, 1990 and 2016,” *New England Journal of Medicine*, vol. 379, no. 25, pp. 2429–2437, Dec. 2018, doi: 10.1056/NEJMoa1804492.

[3] R. A.-A. - *et al.*, “The Impact of Artificial Intelligence on Ghanaian Health Worker Training: Opportunities, Challenges, and Ethical Considerations,” *International Journal For Multidisciplinary Research*, vol. 6, no. 1, Jan. 2024, doi: 10.36948/ijfmr.2024.v06i01.12002.